Multimodal Emotion Recognition

Our project utilizes deep learning to create a multimodal system that identifies seven basic emotions from audio signals (Speech Emotion Recognition – SER) and facial images (Facial Expression Recognition – FER). We separately trained multiple deep learning models for each modality with the same label schema and present comparative evaluation metrics. The models were integrated with a web application interface for real-time prediction.

Rakshith Reddy Dargula - 50591661 - rdargula

Nischith Adavala - 50591641 - nadavala

Sumanth Yalamanchili - 50604274

Project Overview

Speech Emotion Recognition

Analyzes audio signals to detect emotional states using deep learning models trained on benchmark datasets.

Facial Expression Recognition

Processes facial images to classify emotions using CNN architectures trained on standardized datasets.

Multimodal Integration

Combines both modalities to enhance accuracy and robustness of emotion detection for real-world applications.

Dataset	Description	Usability	Notes
CREMA-D	7,442 clips from 91 actors, various ethnicities and emotions	8.75	Balanced across gender, emotions
RAVDESS	1,440 speech files from 24 actors, 8 emotions	8.75	High-quality WAV files
SAVEE	480 files from 4 male English speakers	8.75	Needs female speaker balancing
TESS	2,800 clips from 2 female speakers	8.75	Balances SAVEE

Datasets Used

Dataset	Description	Usability
CREMA-D	7,442 clips from 91 actors, various ethnicities and emotions	8.75
RAVDESS	1,440 speech files from 24 actors, 8 emotions	8.75
SAVEE	480 files from 4 male English speakers	8.75
TESS	2,800 clips from 2 female speakers	8.75
FER2013	35,887 grayscale images (48x48) labeled with 7 emotions	-

We used five benchmark datasets to ensure diversity, speaker/gender balance, and generalization. CREMA-D and RAVDESS provide balanced data across gender and emotions, while SAVEE and TESS complement each other with male and female speakers respectively. FER2013 was used specifically for the facial recognition pipeline.

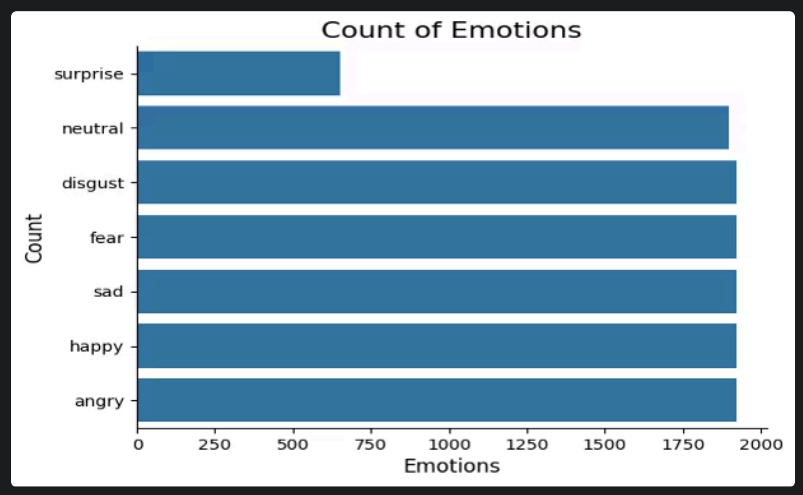
Emotion Label Mapping

Unified 7-Class Schema

All datasets were unified to follow a consistent 7-class schema to ensure standardized training and evaluation across modalities.

Emotion Categories

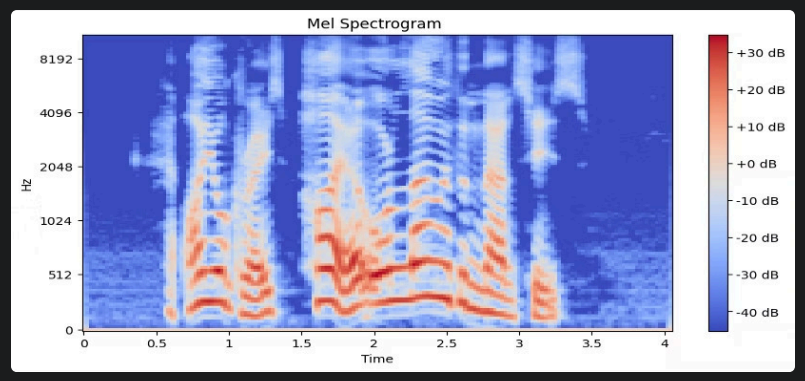
The seven emotion categories used were: neutral, happy, sad, angry, fear, disgust, and surprise.



Label Consistency

Careful manual inspection and relabeling was required to map different naming conventions from various datasets into our unified structure.

This standardized approach allowed us to train models that could recognize the same emotional states regardless of whether they were detected through speech or facial expressions, enabling more consistent multimodal integration.



Audio Data Preprocessing

Data Acquisition

We downloaded datasets using kagglehub.dataset_download() from official Kaggle links. The datasets were loaded into memory using librosa for audio signal processing.

Audio Loading

Each .wav file was loaded using librosa.load(), with a fixed sampling rate of 22050 Hz. We applied padding where audio duration was short and skipped corrupted files.

Feature Extraction

We extracted MFCCs, Zero Crossing Rate (ZCR), and Root Mean Square Energy (RMSE) using Librosa. Features were standardized using StandardScaler, and class labels were encoded with LabelEncoder.

Data Splitting

An 80/20 split of the data into training and test subsets was obtained to ensure proper model evaluation.

Video Data Preprocessing

Dataset Loading

The CSV-formatted FER2013 dataset was loaded using pandas.read_csv().

Image Conversion

Each pixel column was converted into a NumPy array and reshaped into 48x48 images.

Normalization

Pixel values were scaled from [0, 255] to [0, 1] by dividing by 255.

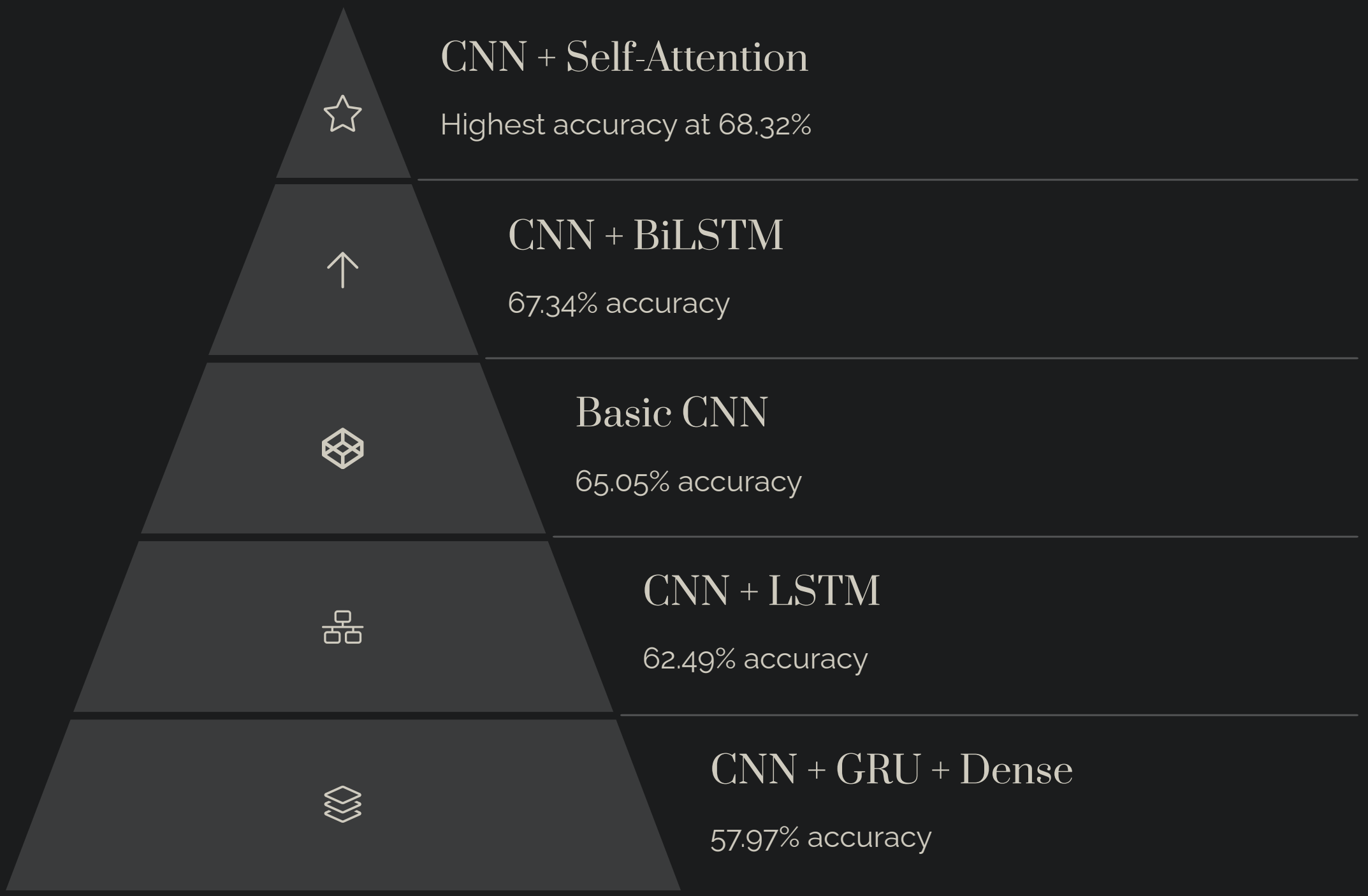
Reshaping

Input image shapes were reshaped from (48, 48) to (48, 48, 1) to match CNN architecture requirements.

Data Splitting

Data was split using train_test_split() with a 70/15/15 ratio (Train/Val/Test).

Speech Emotion Recognition Models



We extracted Mel-Frequency Cepstral Coefficients (MFCCs) from four standard audio emotion datasets, resulting in over 10,000 labeled audio clips. Each model used Adam optimizer, Categorical Cross-Entropy loss function, and employed Early Stopping and ReduceLROnPlateau for regularization and convergence.

Facial Emotion Recognition Models

ResNet-50 (Transfer Learning)

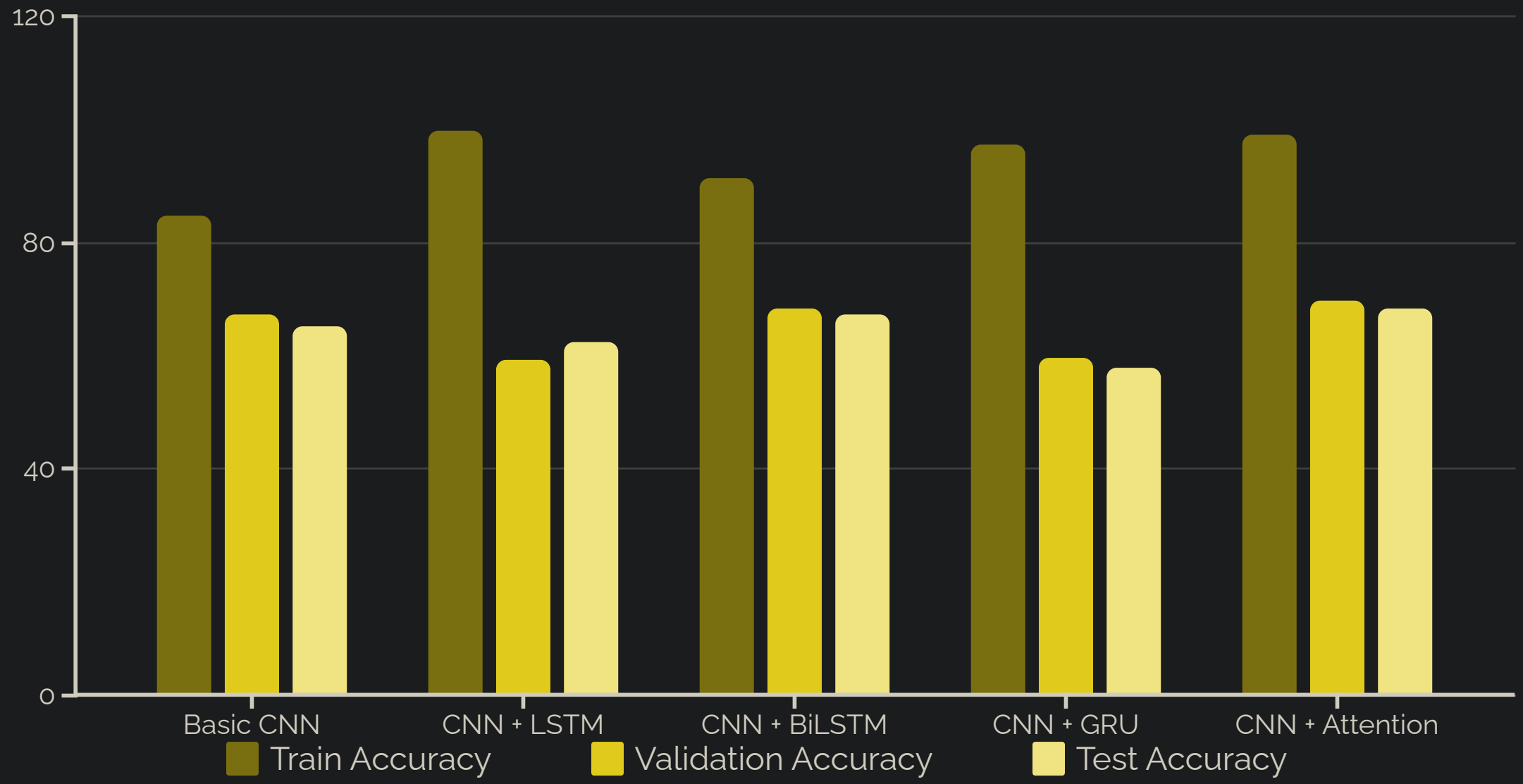
We used pre-trained ResNet50 from Keras with ImageNet weights and frozen base layers. The custom head included GlobalAveragePooling followed by Dense layers. This approach achieved approximately 71% validation accuracy, demonstrating the power of transfer learning for facial emotion recognition.

Custom CNN for FER

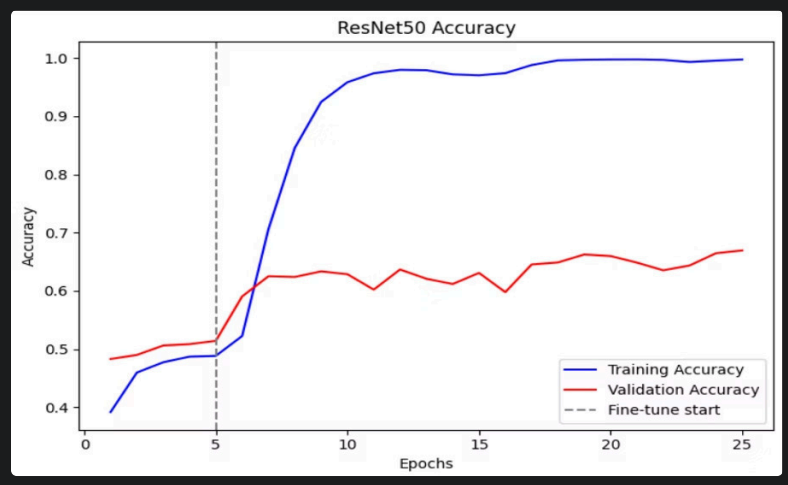
Our custom CNN architecture featured Conv2D stacks with increasing filters (64 → 128 → 256), followed by Flatten, Dropout, and Dense layers. We optimized performance using data augmentation techniques including horizontal flips and brightness adjustments. This model reached approximately 65% accuracy.

Both models were trained on the FER2013 dataset, consisting of 48x48 grayscale facial images labeled with 7 emotion categories. The superior performance of the ResNet-50 model validates the importance of transfer learning in facial expression recognition tasks.

Model Performance Summary



The chart illustrates the performance metrics of our five SER models. The CNN + Self-Attention model achieved the highest test accuracy at 68.32%, followed by CNN + BiLSTM at 67.34%. Note the significant gap between training and test accuracies in most models, indicating some degree of overfitting despite regularization techniques.



Facial Recognition Model Performance

99.72%

ResNet-50 Training Accuracy

Shows excellent learning on training data

66.94%

ResNet-50 Test Accuracy

Indicates good generalization

62.38%

Custom CNN Training Accuracy

Lower but more consistent with test results

64.94%

Custom CNN Test Accuracy

Shows less overfitting than ResNet-50

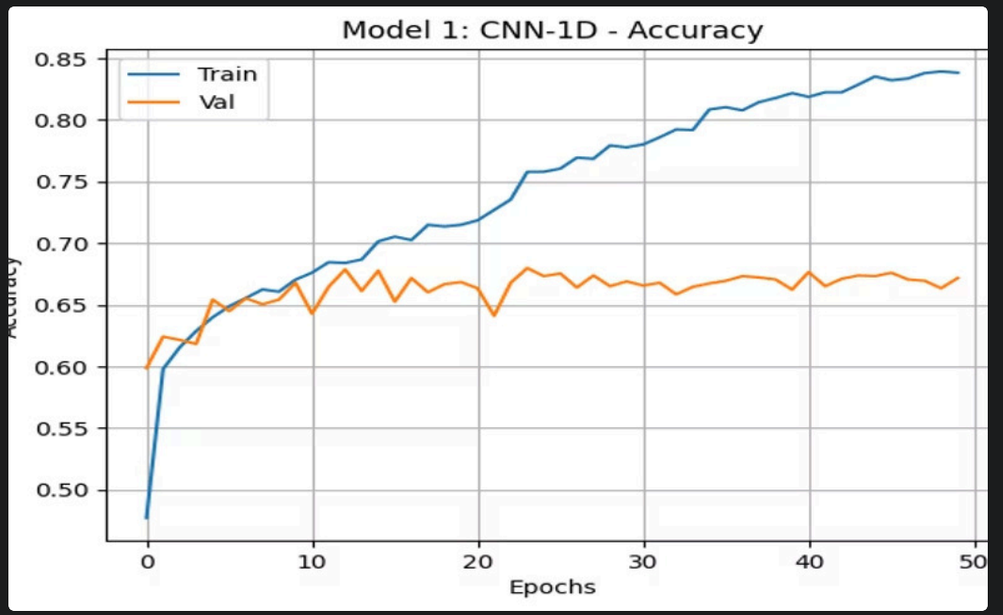
The ResNet-50 model demonstrates superior performance compared to our Custom CNN architecture, though with more pronounced overfitting. The Custom CNN shows more consistent performance between training and testing phases, suggesting better generalization despite lower overall accuracy.

Key Evaluation Metrics

Model Name	Precision	Recall	F1 Score	Accuracy
Basic CNN	0.66	0.65	0.65	0.65
CNN + LSTM	0.65	0.62	0.62	0.62
CNN + BiLSTM	0.68	0.67	0.67	0.67
CNN + GRU + Dense	0.59	0.58	0.58	0.58
CNN + Self-Attention	0.69	0.68	0.68	0.68
ResNet-50	0.685	0.653	0.665	0.669
Custom CNN	0.614	0.653	0.623	0.649

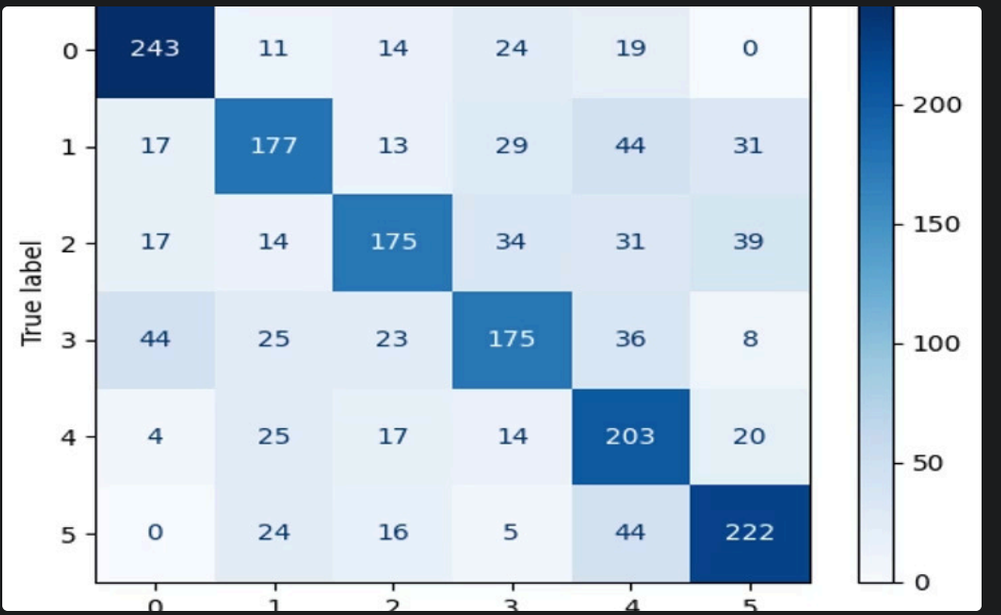
This comprehensive evaluation shows that CNN + Self-Attention achieved the best overall performance for audio with an F1-score of 0.68, while ResNet-50 led the visual models with an F1-score of 0.665. The metrics reveal that precision was strongest for the "happy" class across models, while "fear" and "disgust" classes showed the weakest recall in both modalities.

Basic CNN Performance



Accuracy Plot

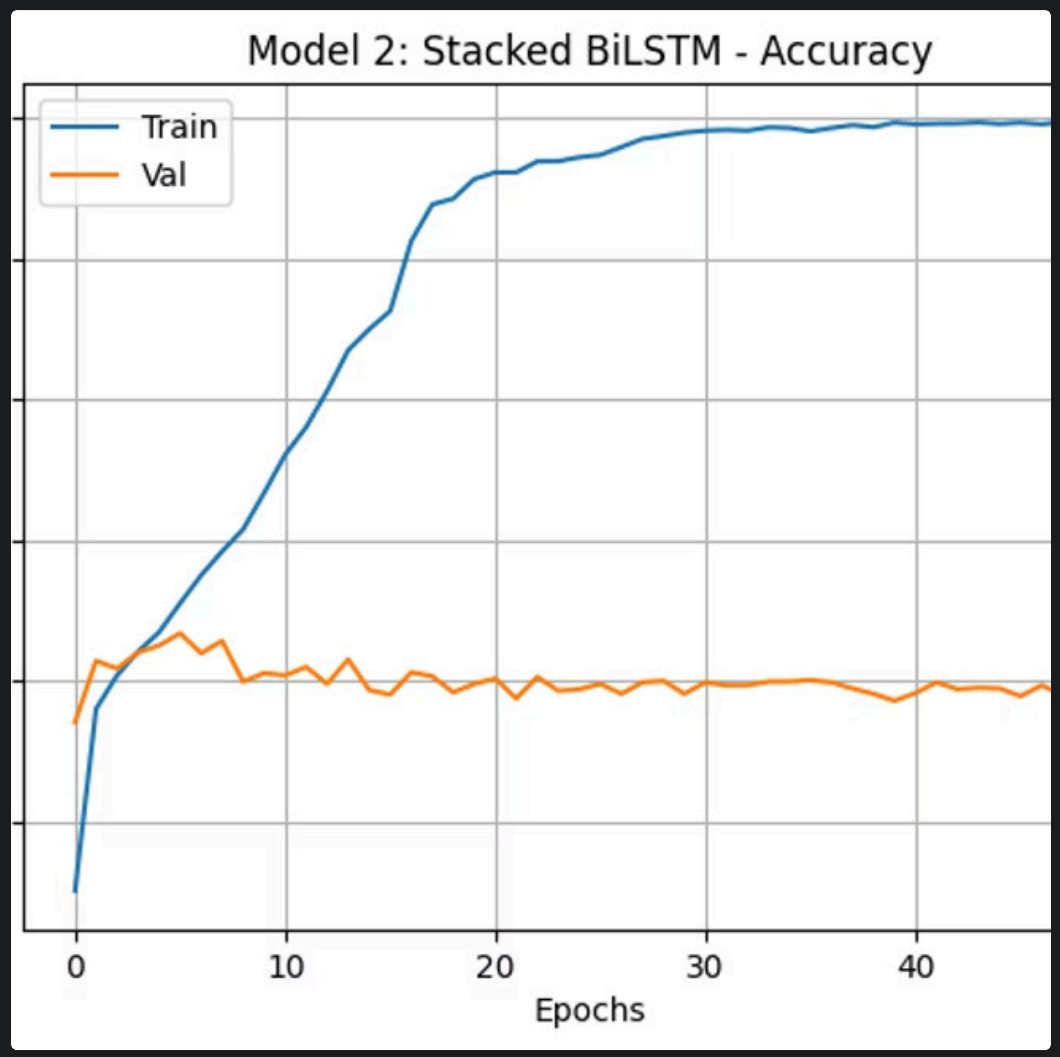
The training accuracy rises steadily to around 85% while validation accuracy plateaus near 67%, indicating some overfitting despite regularization techniques. The gap between training and validation curves suggests room for improved generalization.

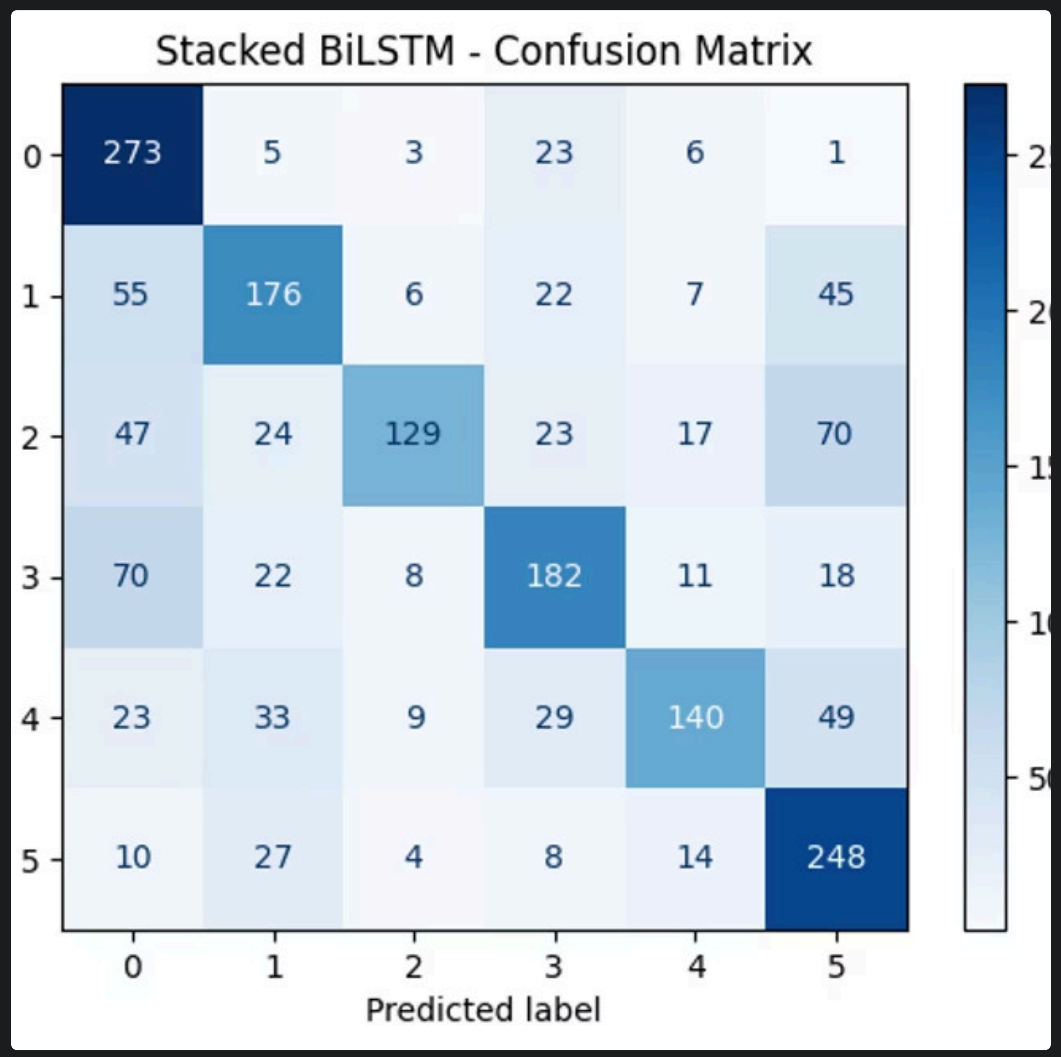


Confusion Matrix

The confusion matrix reveals stronger performance on "happy" and "neutral" emotions, with more confusion between similar emotion pairs like "sad"/"fear" and "angry"/"disgust". This pattern is consistent with human perception challenges for these emotion pairs.

CNN + LSTM Performance





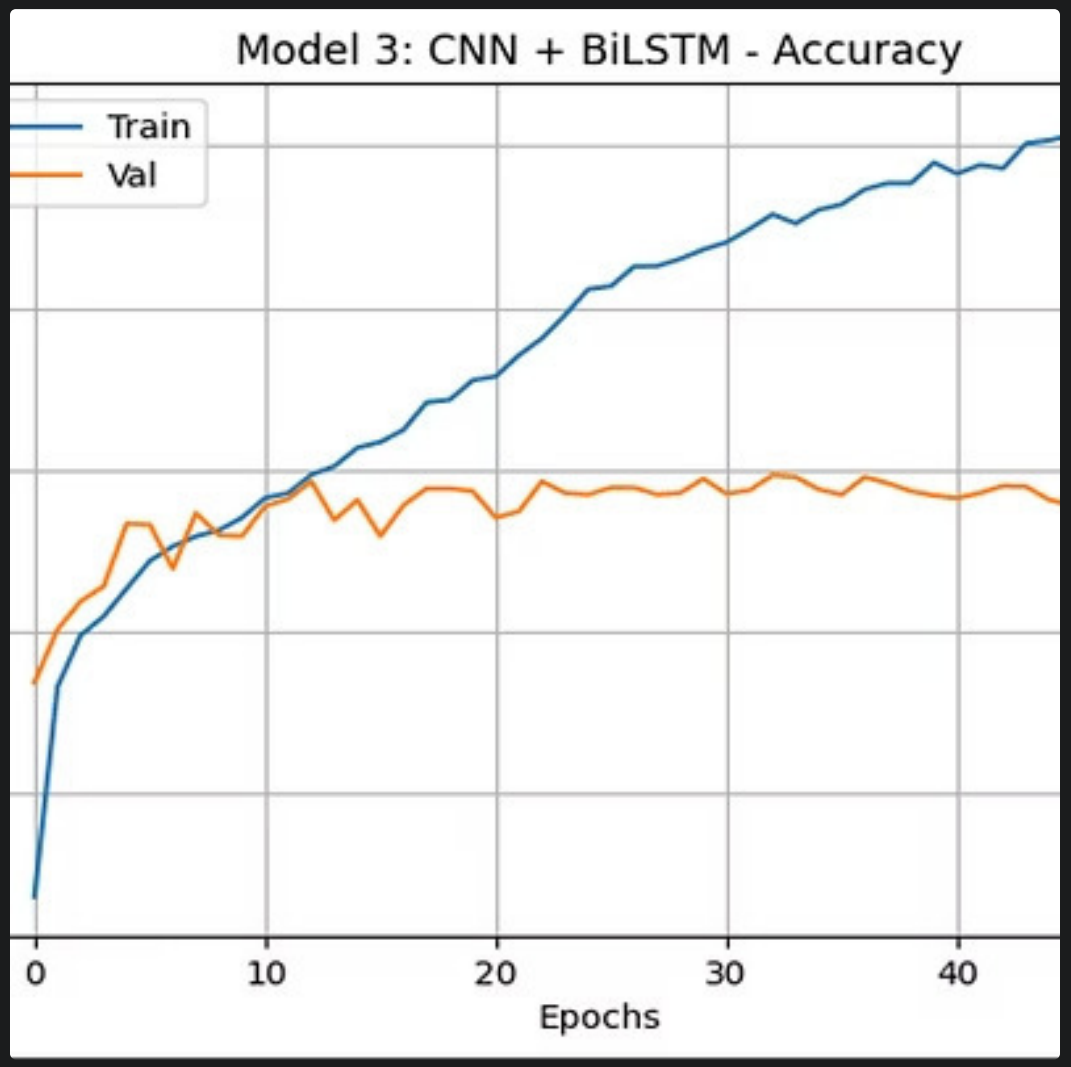
Accuracy Plot

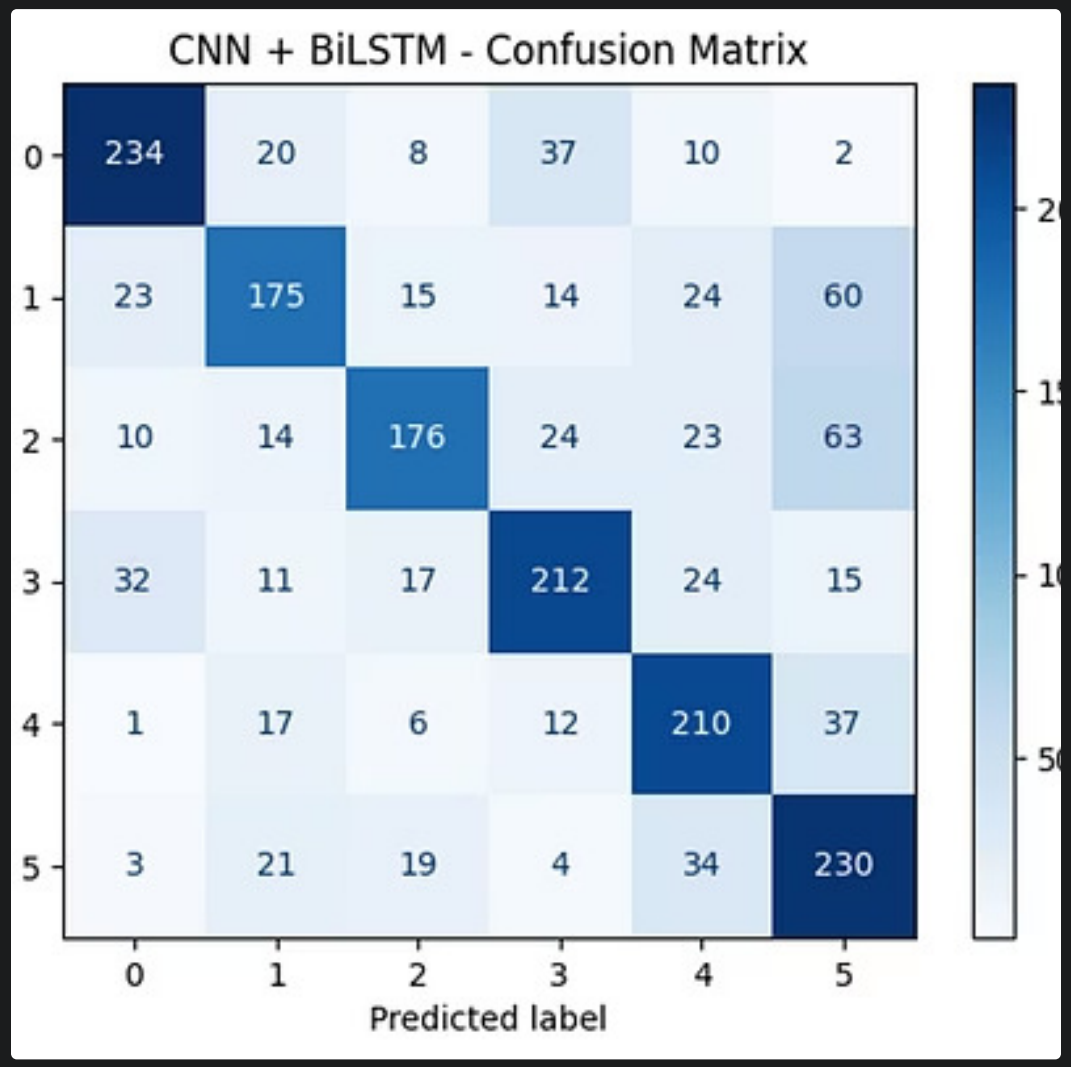
The CNN + LSTM model shows significant overfitting with training accuracy approaching 100% while validation accuracy remains around 59%. This suggests the model memorized training data rather than learning generalizable patterns despite dropout layers.

Confusion Matrix

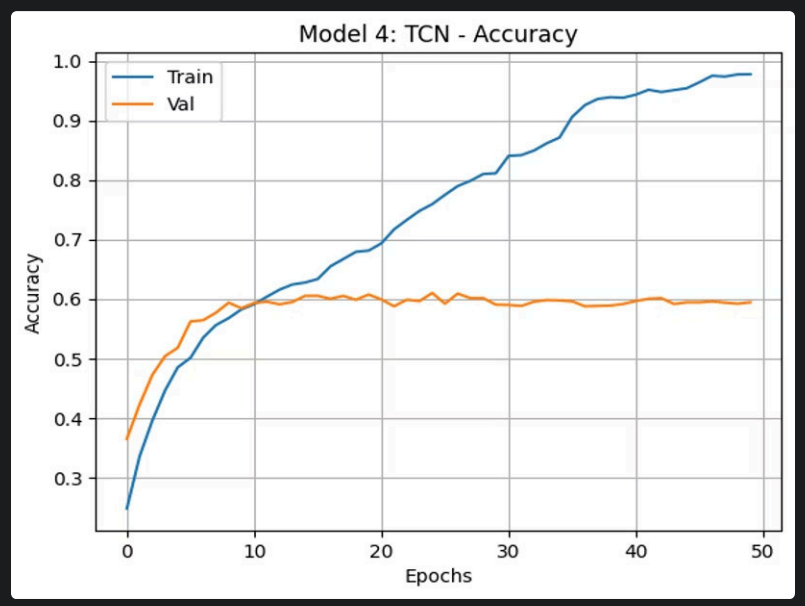
The confusion matrix shows improved detection of temporal emotion patterns compared to the basic CNN, particularly for "surprise" and "angry" emotions. However, it struggles with distinguishing between "fear" and "sad" emotions, which share similar acoustic properties.

CNN + BiLSTM Performance





The CNN + BiLSTM model shows improved performance over the standard LSTM variant, with better balance between training (91.19%) and validation accuracy (68.41%). The bidirectional approach captures both forward and backward context in audio features, resulting in better performance on emotions like "disgust" and "fear" that rely on tone shifts. The confusion matrix shows reduced misclassifications between similar emotion pairs.



CNN + GRU + Dense Performance

Training Pattern

Training accuracy reached 97.43% while validation accuracy peaked at 59.42%, showing significant overfitting despite using GRU units which are typically more resistant to this problem.

Model Architecture

Replaced LSTM with GRU for lighter computation while maintaining similar theoretical capabilities. Added dense layers with dropout to combat overfitting, though with limited success.

Classification Results

Showed the weakest overall performance at 57.97% test accuracy, with particular difficulty distinguishing between "neutral" and "sad" emotions.

CNN + Self-Attention Performance

Best Performer

Achieved highest test accuracy of 68.32%

Attention Mechanism

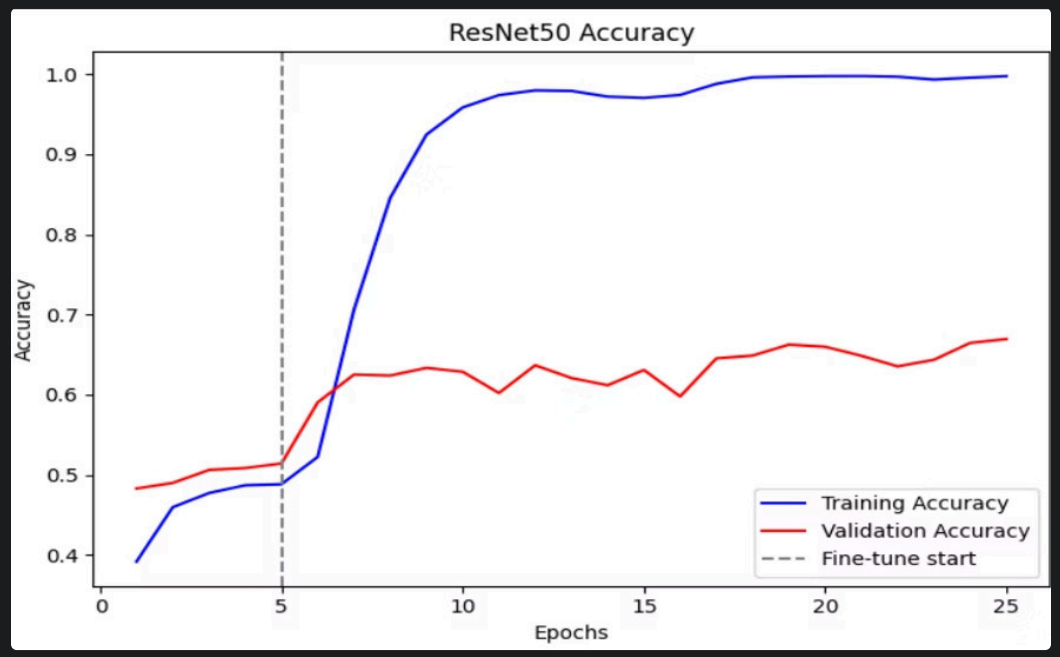
Incorporated self-attention to emphasize important MFCC regions

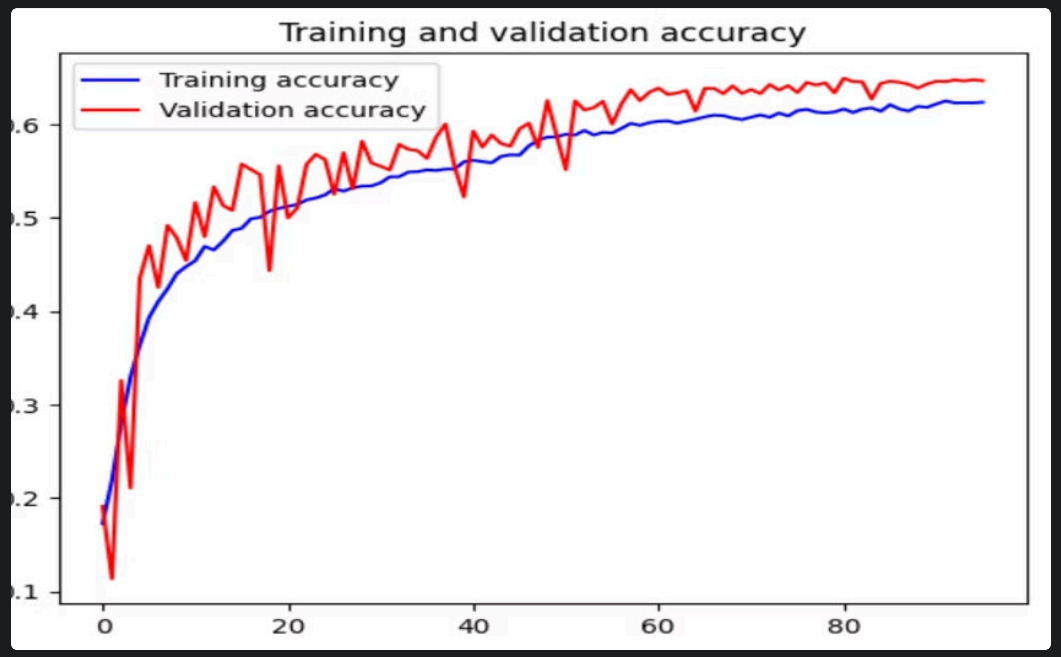
Balanced Learning

Better performance across all emotion classes

The CNN + Self-Attention model represents our most sophisticated architecture for speech emotion recognition. By incorporating an attention mechanism, the model can focus on the most emotionally salient parts of the audio signal. This approach achieved the highest test accuracy of 68.32% and showed more balanced performance across all emotion classes, particularly improving on traditionally difficult emotions like "fear" and "disgust".

ResNet-50 & Custom CNN Performance





ResNet-50 Results

The ResNet-50 model leveraged transfer learning to achieve 66.94% test accuracy. The confusion matrix shows strong performance on "happy" emotions but struggles with "fear" and "disgust" classes. The training curve reveals significant overfitting despite using pre-trained weights.

Custom CNN Results

Our custom CNN architecture reached 64.94% test accuracy with less overfitting than ResNet-50. The model shows more balanced performance across emotion classes but with lower overall accuracy. Data augmentation techniques helped improve generalization.

Challenges Faced

Dataset Integration

Different formats and label naming conventions required careful manual inspection and relabeling

Computational Limits

GPU constraints, session timeouts, and memory limits slowed experimentation

Data Imbalance

Underrepresented emotions like disgust, fear, and surprise led to biased predictions

Audio Processing

WAV files failed due to encoding issues, length inconsistencies, and sampling rate mismatches

Overfitting

Deep models showed high training accuracy but poor generalization despite regularization

Future Use Cases and Extensions

Real-Time Emotion Monitoring

Deploy trained models using web or mobile interfaces for live microphone input and webcam streams, enabling continuous affect tracking during calls or video chats.

Speech-to-Text-Based Detection

Include textual modality using speech-to-text conversion with models like Whisper or Google Speech API, then apply transformer-based models to detect emotion from text content.

Multimodal Fusion

Extend current late-fusion strategy into a fully integrated model that processes audio, visual, and text inputs simultaneously using multimodal transformers or cross-attention architectures.

Emotion-Aware Assistants

Integrate emotion detection into chatbots, customer service systems, and healthcare assistants to identify user states like frustration, satisfaction, stress, or anxiety.

Project Contributions and References

Team Contributions

Team Member	Project Part	Contribution(%)
nadavala	Speech emotion recognition pipeline	33.33%
rdargula	Facial expression recognition models	33.33%
sumanthy	Model integration and reporting	33.33%

Key References

- CREMA-D, RAVDESS, TESS, and SAVEE datasets from Kaggle
- FER2013 Facial Emotion Recognition Dataset
- TensorFlow Documentation
- Librosa Audio Processing Library
- GitHub repositories for multimodal emotion recognition